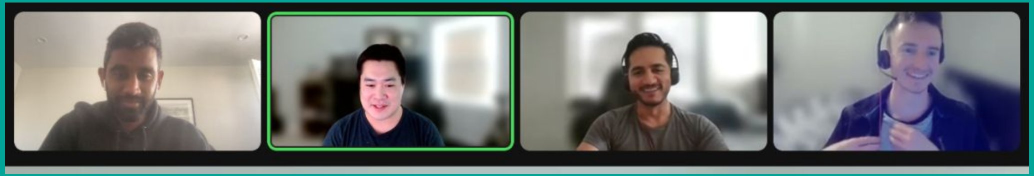


Flight Delay Prediction

Section 4 Group 2

Jumping the Spark

Ahmad Azizi, Evan Chan, Kolby Devery, Chetan Munugala

Berkeley
UNIVERSITY OF CALIFORNIA

Outline

- Business Case
- Dataset
- Feature Engineering
- EDA
- Data Lineage
- Preprocessing
- Modeling and Results
- Gap Analysis
- Limitations/Future Work
- Key Takeaways

Business Case

- Predict departure delays two hours prior to expected departure time.
- A delay is defined as 15 mins or more.
- Airports in the 50 states only
  - No territories
- Using 3 primary datasets
- Leverage ML approaches
- Success Metric: F 0.5 score

Total Cost of Delay in the U.S. (dollars, billion)

	2016	2017	2018	2019
Airlines	5.6	6.4	7.7	8.3
Passengers	13.3	14.8	16.4	18.1
Lost Demand	1.8	2.0	2.2	2.4
Indirect	3.0	3.4	3.9	4.2
Total	**23.7**	**26.6**	**30.2**	**33.0**

https://www.faa.gov/data_research/aviation_data_statistics/media/cost_delay_estimates.pdf

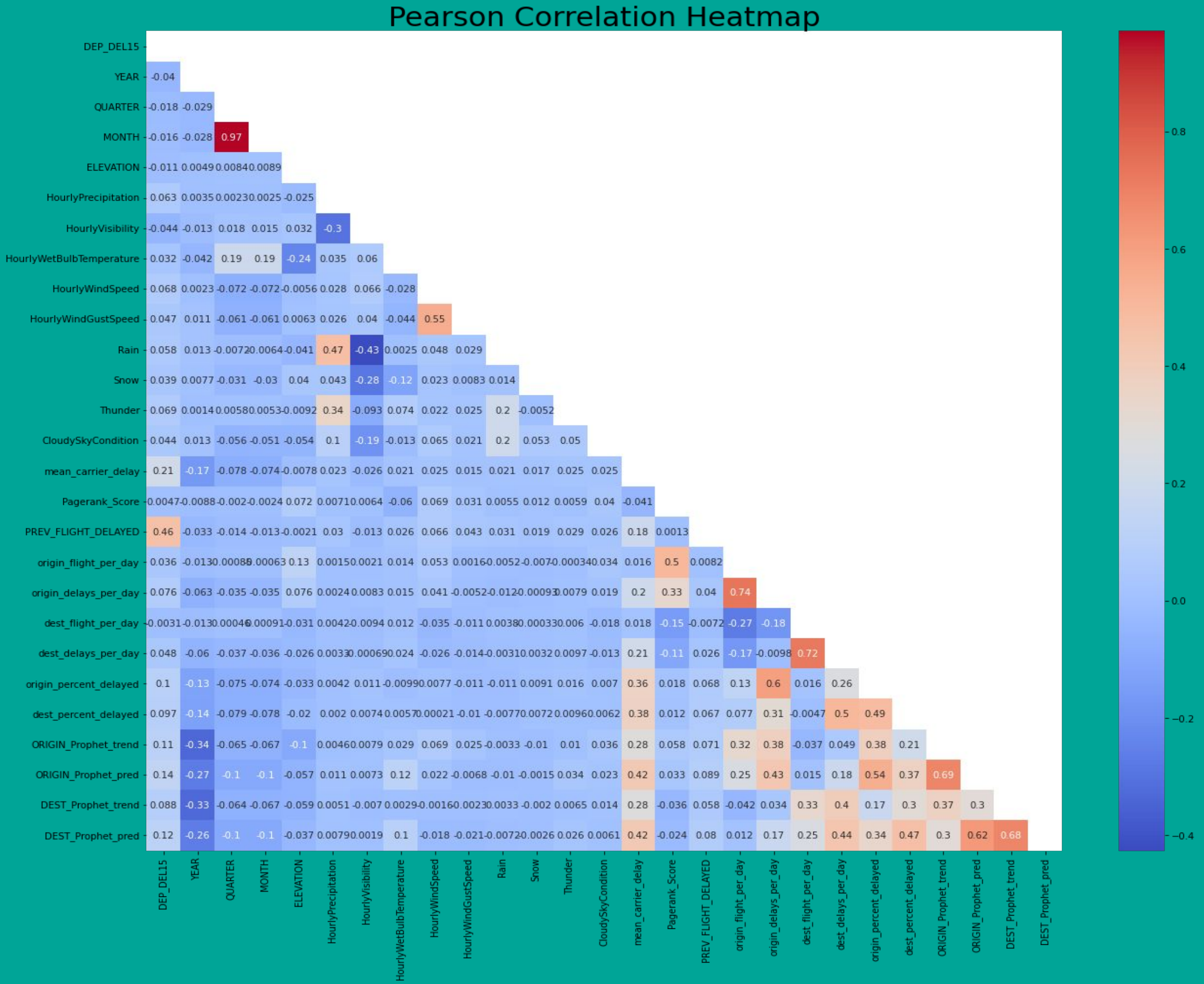
Dataset Overview

- Primary Datasets
  - US Flights (2015-2021) dataset from the US Department of Transportation
  - Weather stations dataset from the US Department of Transportation
  - Weather dataset from the National Oceanic and Atmospheric Administration Repository
- Secondary Dataset
  - Airports dataset from https://openflights.org/

Feature Engineering

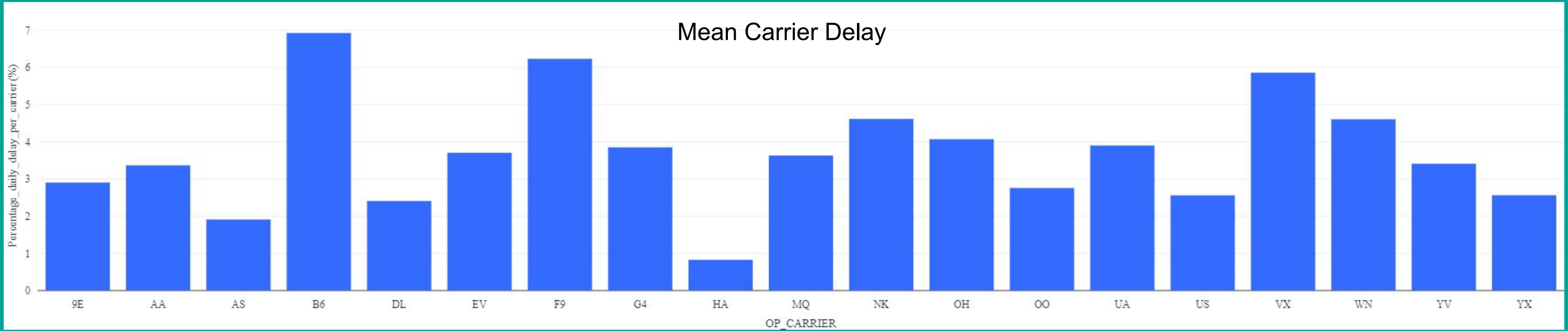
- Text based features
  - Weather condition text columns
    - HourlyPresentWeatherType Codes
    - HourlySkyConditions Codes
- Graph Based Features
  - Pagerank of Airports as Nodes
- Frequency Related Features with Time Component
  - Flag for holiday period
  - Previous flight delayed by TAIL_NUM
  - Mean carrier delay for the previous day
  - Number of flights and delays by airport for the prior two days.
- Time Series Features
  - Percent flights delayed for the prior two days.
  - Prophet forecast and trend for percent flight delayed

EDA


Pearson Correlation Heatmap
DEP_DEL15
YEAR
QUARTER
MONTH
ELEVATION
HourlyPrecipitation
HourlyVisibility
HourlyWetBulbTemperature
HourlyWindSpeed
HourlyWindGustSpeed
Rain
Snow
Thunder
CloudySkyCondition
mean_carrier_delay
Pagerank_Score
PREV_FLIGHT_DELAYED
origin_flight_per_day
origin_delays_per_day
dest_flight_per_day
dest_delays_per_day
origin_percent_delayed
dest_percent_delayed
ORIGIN_Prophet_trend
ORIGIN_Prophet_pred
DEST_Prophet_trend
DEST_Prophet_pred
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

EDA

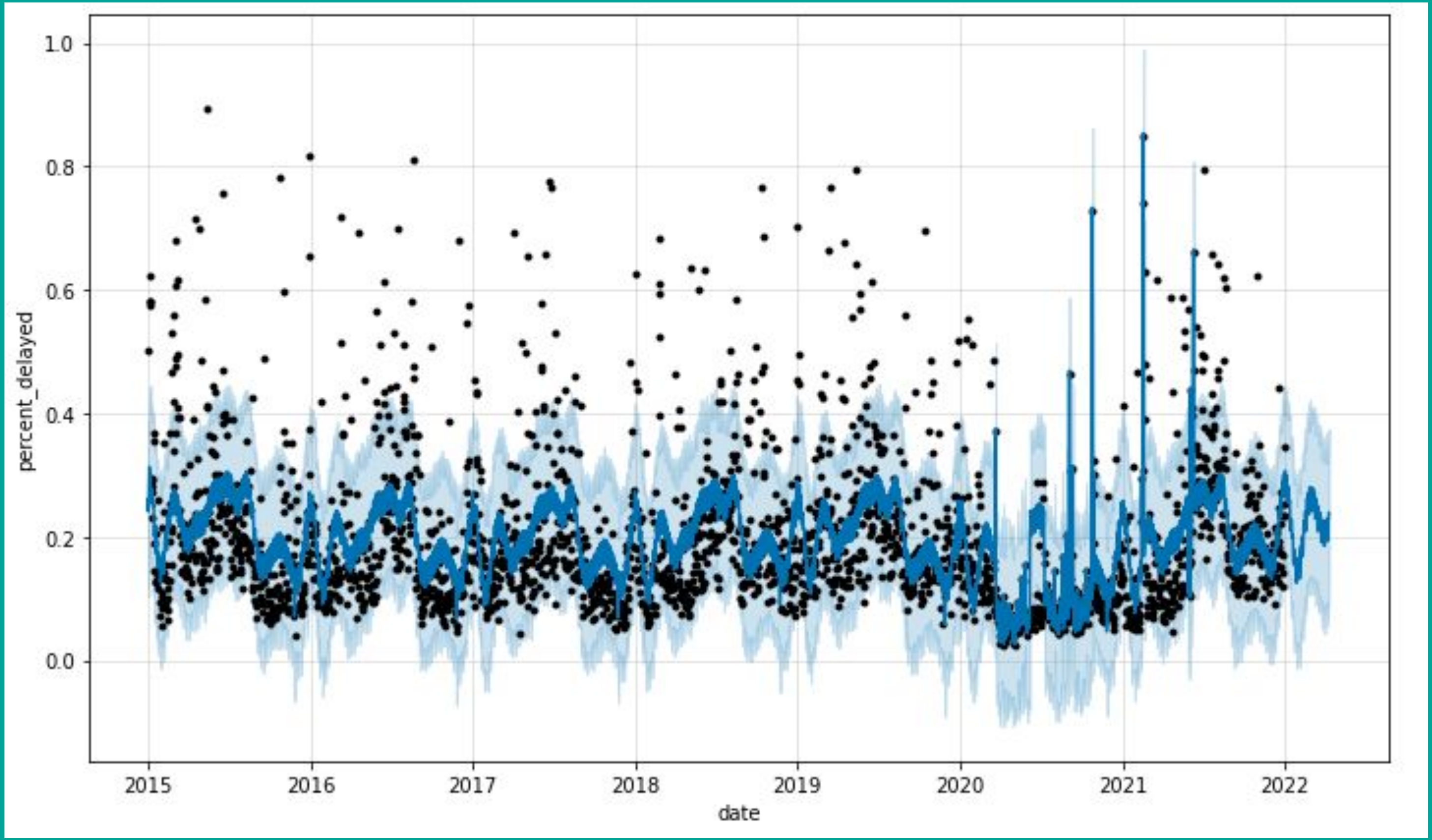
Page Rank Scores
count
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
OMA MSP LGA ATL MYR MCO ALB SFB ATW DEN DFW MKE SDF TYS SFO PIE ORD IND CLT FLL RAP XNA PHL ABE GRR LAS MDT DTW PDX AZA LGB BIS LAX EWR ACV CHA ANC PLN APN SLC OAK PVU JMS AUS DVL IAH PHX MDW LFT GEG
PageRank Score



Mean Carrier Delay
Percentage_daily_delay_per_carrier(%)
7
6
5
4
3
2
1
0
9E AA AS B6 DL EV F9 G4 HA MQ NK OH OO UA US VX WN YV YX
OP_CARRIER

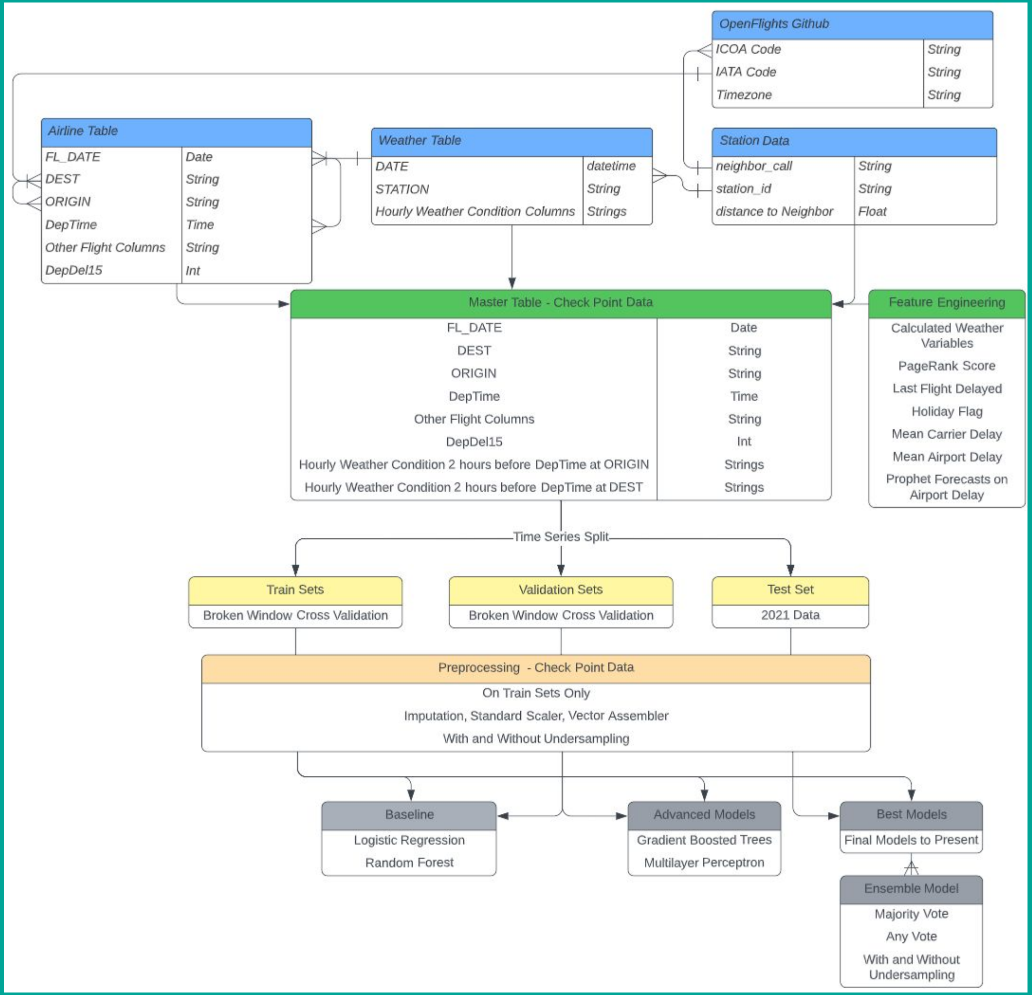
EDA

Delay Forecast Using Prophet (Dallas/Fort Worth Airport)



Data Lineage

- Join Data
- Engineer Features
- Time Series Split
  - 2021 held out for blind test
  - 2015-2020 for cross-validation



Preprocessing

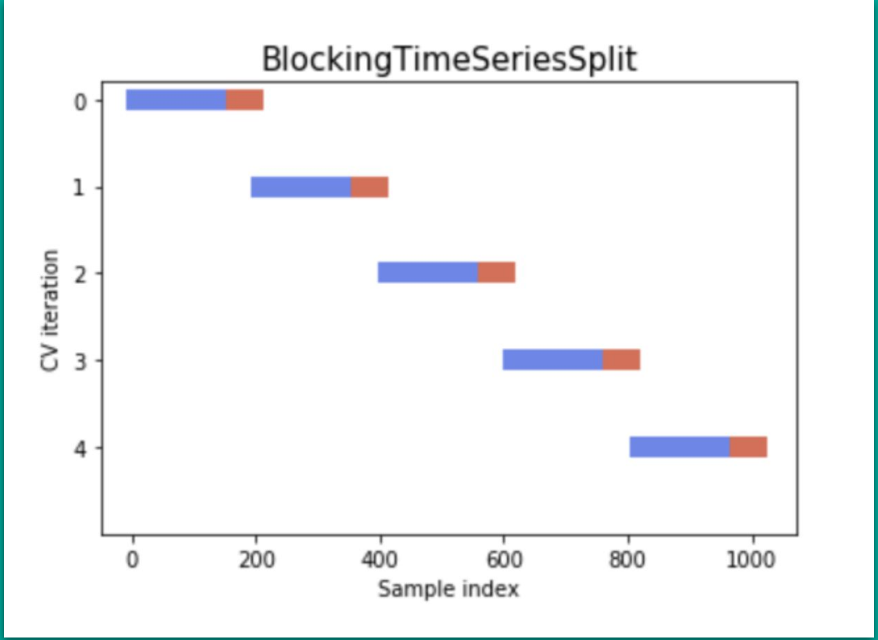
- For Cross-Validation process by fold, same for full train and test sets
  - Imputation
  - Scaling
  - Vector Assembler
  - 34 minutes to process CV folds

- Saving Cleaned Tables and Preprocessed Data crucial for fast execution
  - Time series split means that the folds are always the same and don't have to be reprocessed for each model
  - Cross Validation on saved folds, test scores on full train and test sets

Metrics and Cross Validation

- Evaluation Metric:
    - F0.5 Score

- Broken Window Cross Validation
    - 5 folds over 6 years
    - Avoid validating on same time for each calendar year



https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4

Models

- Logistic Regression (Baseline)
- Logistic Regression
- Random Forest
- Gradient Boosting Classifier
- Neural Network

- NOTE: Class imbalance!

Logistic Regression

- Hyperparameters:
  - Threshold: [0.3, 0.5, 0.8]
  - regParam: [0.01, 0.1, 0.5, 1.0, 2.0]
  - elasticNetParam: [0.0, 0.25, 0.5, 0.75, 1.0]
  - maxIter: [1, 5, 10, 20, 50]

Class Imbalance Handling	Optimal Hyperparameters	F0.5 Score	Execution Time	Computation Resource
None	Threshold: 0.5, regParam: 0.01 elasticNetParam: 1 maxIter: 5	0.59	2.36 minutes	5 workers, 4 cores each

Random Forest

- Hyperparameters:
  - maxDepth: [5, 10]
  - numTrees: [32, 64, 128]

Class Imbalance Handling	Optimal Hyperparameters	F0.5 Score	Execution Time	Computation Resource
None	maxDepth: 10 numTrees: 32	0.59	13.56 minutes	10 workers, 4 cores each

Gradient Boosted Classifier

- Hyperparameters:
  - maxDepth: [5,10]
  - minInfoGain: [0.0, 0.2, 0.4]
  - maxBins: [32,64]
  - Undersampling: [True,False]

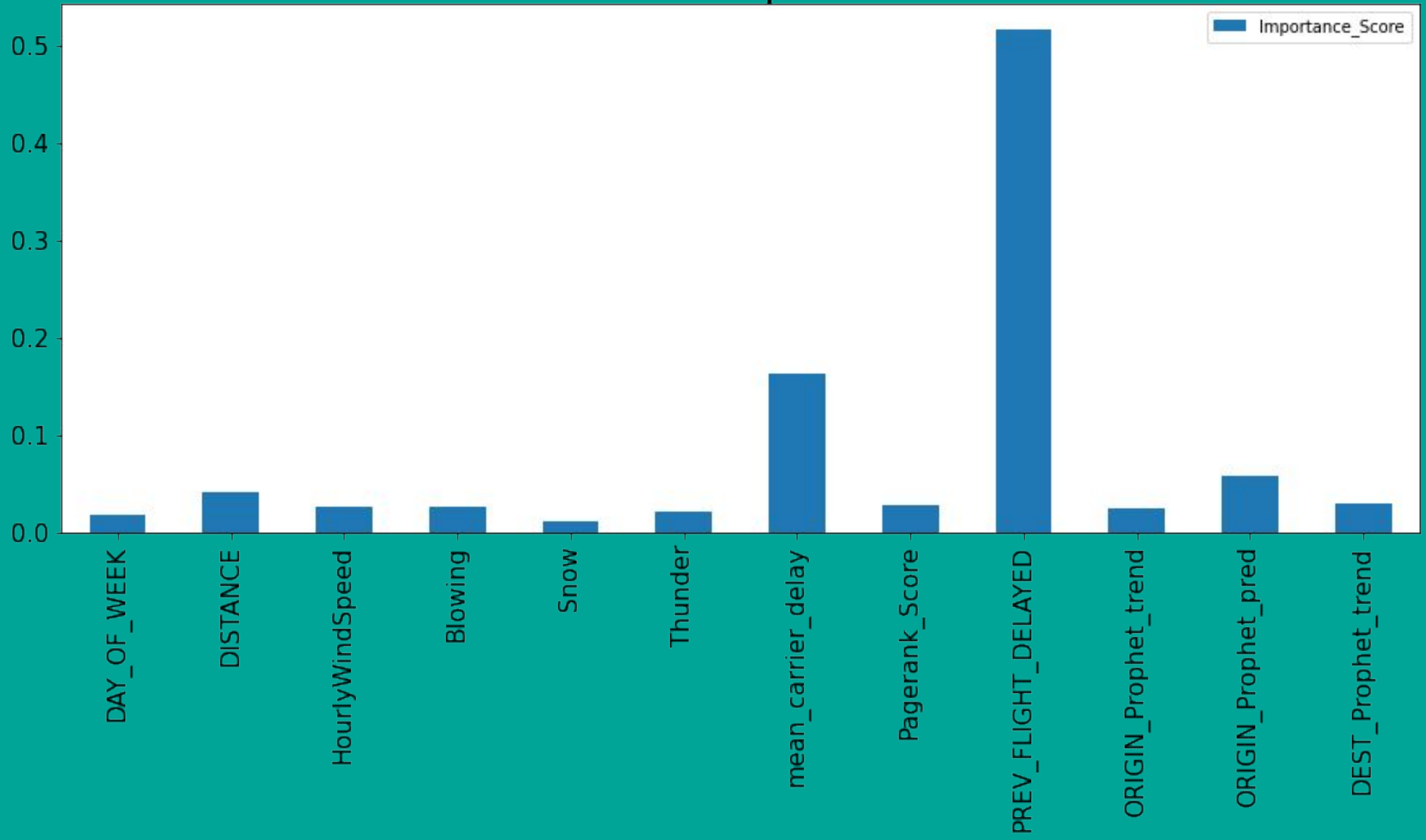
Class Imbalance Handling	Optimal Hyperparameters	F0.5 Score	Execution Time	Computation Resource
None	maxDepth: 5 minInfoGain: 0 maxBins: 64	0.59	19.12 minutes	5 workers, 4 cores each

Multilayer Perceptron

- Hyperparameters:
  - MaxIter: [50,100,200]
  - Layers: [[2,26,38],[2,26,26,38]]
  - BlockSize: [32,64]
  - Solver: [‘gd’,’l-bfgs’]
  - Undersampling: [True,False]

Class Imbalance Handling	Optimal Hyperparameters	F0.5 Score	Execution Time	Computation Resource
None	MaxIter: 100 Layers: [2,26,38] BlockSize: 64 Solver: ‘l-bfgs’	0.59	34 minutes	10 workers, 4 cores each

Feature Importance



Key Takeaways

- All models hit ceiling of F.5 = ~.59
- Increased model performance would likely require further feature engineering

Model	Class Imbalance Handling	Test Data F.5 Score	HyperParameters	Execution Time	Computation Resources
Logistic Regression	None	0.59	threshold=0.3, regParam=0.01, elasticNetParam=1.0, maxIter=5	2.36 minutes	5 workers, 20 cores
Logistic Regression	Undersampling	0.56	threshold=0.5, regParam=0.01, elasticNetParam=1.0, maxIter=5	24.83 seconds	10 workers, 40 cores
Random Forest	None	0.59	maxDepth=10, numTrees=32	13.56 minutes	10 workers, 40 cores
Random Forest	Undersampling	0.57	maxDepth=10, numTrees=128	10.25 minutes	10 workers, 40 cores
GBT	None	0.59	maxDepth=5, minInfoGain=0, maxBins=64	19.12 minutes	10 workers, 40 cores
GBT	Undersampling	0.56	maxDepth=5, minInfoGain=0, maxBins=64	14.88 minutes	10 workers, 40 cores
MLPC	None	0.59	maxIter=100, layers=[39,26,2], blockSize=64, solver='l-bfgs'	34.69 minutes	10 workers, 40 cores
MLPC	Undersampling	0.55	maxIter=100, layers=[39,26,2], blockSize=64, solver='l-bfgs'	19.27 minutes	10 workers, 40 cores

Gap Analysis

- Few groups chose the same evaluation metric
- Difficult ‘apples to apples’ comparison
- Very competitive F.5 score

Limitations and Future Work

- Join is missing several hundred thousand rows
- Pull in 2022 data
- Explore further methods of over/undersampling
- Explore other PageRank techniques
- Explore other Prophet/forecasting approaches
- Further feature engineering

Questions?